Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 1 | Jan 2017

A STUDY ON TRAVEL TIME ESTIMATION FOR TAXI TRIPS

Niranjan Joshi [1] | Manmat Hotalappa [1] | Kajal Gadade [1]

[1] Department of Computer Engineering, NBN Sinhgad School of Engineering, Pune,411041.

ABSTRACT

With companies like Uber, OLA, request and dispatch of taxi has become easier and accessible to a wide range of people. The problem of being able to automatically assign the taxi for a given trip request is important from both the service time and economics perspective of this business. One of the important aspect of this decision is to determine which near by taxi can be available early and hence predicting trip times becomes important to understand. In this problem, we aim at using machine learning techniques to predict trip travel times based on the characteristics of the trip. Kaggle has provided the data on trip times of various taxi trips as a part of it's competition. This data can be used to both train and test the models.

1. INTRODUCTION

With comipanies like Uber, OLA, request and dispatch of taxi has become easier and accessible to a wide range of people. The problem of being able to automatically assign the taxi for a given trip request is important from both the service time and economics perspective of this business. One of the important aspect of this decision is to determine which near-by taxi can be available early and hence predicting trip times becomes important to understand.

The travel time of the trip depends on several factors such as origin, destination, the path/trajectory from origin to destination and more importantly real-time traffic conditions. Traffic conditions in turn are specific to day and/or time of the travel as well as the travel route. Driver and taxi profiles might also affect the travel time. Some drivers might have tendency to drive faster compared to some others. Some taxis might be well maintained and hence may have good impact of travel time, whereas some other which are poorly maintained might have problems with running at high speed even if the road is empty. Understanding dependency of travel time on these factors is an important study before we actually try to solve the prediction problem here

.

With the advances in machine learning and recent advances in deep neural networks, it is becoming easier to build prediction models that are robust and yet can perform in realtime Though training these models is still computationally intensive, using built prediction model in real-time is becoming common to perform prediction tasks. In this problem, we aimat using machine learning techniques to predict trip travel times based on the characteristics of the trip

The competition to solve similar problem was hosted by [5] in affiliation to ECML/PKDD in year 2015-16. The competition was based on Taxi trip travels in Porto in the year 2014-15. The dataset provided as a part of the competition has information of approximately 1.7 billion trips. Several contestants participated and have developed their models to address the problem. We present here the study of already developed models and the scope of improvement to come up with better results. We plan to take on this challenge to further improve the model by addressing some of the aspects that are not covered by existing models.

The competition hosted by [5] consisted of two problems The first one was to predict the destination of a given taxi trip based on the partial trajectory and other trip information. The second one which we are trying to solve here was to predict the travel time of a given taxi trip based on the starting time,

partial trajectory and other trip information such as day of the trip, taxi used etc. Next we describe the dataset that is made dataset available as a part of this competition. Training data containing the following parameters for each taxi trip

- TRIP ID: (String) It contains an unique identifier for each trip;
- CALL TYPE: (char) It identifies the way used to demand this service. It may contain one of three possible values:
- A if this trip was dispatched from the central;
- B if this trip was demanded directly to a taxi driver on a specific stand;
- C otherwise (i.e. a trip demanded on a random street).
- ORIGIN CALL: (integer) It contains an unique identifier for each phone number which was used to demand, at least, one service It identifies the trips customer if CALL TYPE=A. Otherwise, it assumes a NULL value;
- ORIGIN STAND: (integer): It contains an unique identifier for the taxi stand. It identifies the starting point of the trip if CALL TYPE=B. Otherwise, it assumes a NULL value;
- TAXI ID: (integer): It contains an unique identifier for the taxi driver that performed each trip;
- TIMESTAMP: (integer) Unix Timestamp (in seconds). It identifies the trips start;
- DAYTYPE: (char) It identifies the day type of the trips start. It assumes one of three possible values

:

- B if this trip started on a holiday or any other special day (i.e. extending holidays, floating holidays, etc.);
- C if the trip started on a day before a type-B day;
- A otherwise (i.e. a normal day, workday or weekend).
- MISSING DATA: (Boolean) It is FALSE when the GPS data stream is complete and TRUE whenever one (or more) locations are missing
- POLYLINE: (String): It contains a list of GPS coordinates(i.e. WGS84 format) mapped as a string. The beginning and the end of the string are identified with brackets(i.e. [and], respectively). Each pair of coordinates is also identified by the same brackets as [LONGITUDE, LATITUDE]. This list contains one pair of coordinates for each 15 seconds of trip. The last list m corresponds to the trips destination while the first one represents its start;As mentioned earlier the goal is to predict destination and travel time of the trip. For test data destination prediction, output is of the form (TRIP ID, LATTITUDE, LONGITUDE)For travel-time, it is of the form (TRIP ID, TRAVEL TIME)

II. LITERATURE SURVEY

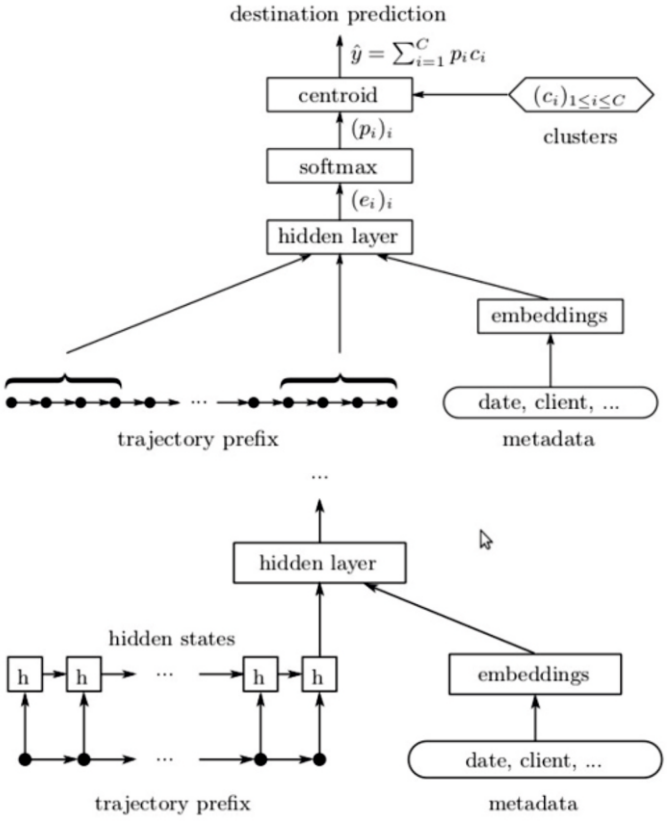
Several teams participated in the competition and reported their model and results achieved using their models. [7]summarizes the competition, dataset, participation and the results. Next, we present here relevant work published through this competition. Since, knowing destination of the trip can also help (in fact, that is a prerequisite) for travel time prediction, we are also including the models used by destination prediction as part of this review.

A. ANN based approach

The winner team for desstination prediction challege, [1]used a model based on Artificial neural networks with multi-layerperceptron architecture (MLP). Figure 1 shows the model. Since, the provided dataset consists of varying length data in POLY LINE field, but MLP requires fixed length input, they used first and last k GPS points in each POLY LINE as input to the model. This gives 2k GPS points or 4k numerical values (longitude and lattitude for each GPS point).The approach also used metadata by creating an embedding for each metadata field (such as day-of-the-week, origin, taxi stand etc). The embedding along with 4k numerical values obtained above form the feature space. Each input (taxi trip) is represented in this feature space. For the destination prediction, the destinations were first grouped into few thousand clusters using mean-shift clustering algo-

rithm. A weighted average of centeres of these clusters was used to predict the final destination. Since, the model didn't train very well using Haversine distance function, a simpler equirectuangular distance was used. Stochastic Gradient descent with momentum was used to minimize mean equirectanguarl distance between
predictions and actual destination points. This was mostly an automated approach and didn't require any manual processing. The network consisted of 500 ReLu neurons

.The paper also describes alternative approaches using Recurrent Neural Network (RNN) and Bi-directional RNN (BRNN) as shown in Figure 2 and Figure 3 respectively.



B. Kernel Regression method

The other team [6] used Kernel Regression (KR) method for solving the problem. This work involved some preprocessing on data especially to figure out missing data in polyline by observing large jumps between two consecutive GPS locations. The typical features used for Kernel Regression method here included full trajectory, last d meter of trajectory, Euclidean and haversine distance between end-points of trajectory, direction of movement (going in-the-city vs going out-of-the-city)This approach also used the contextual features which included day of the week, taxi id, call id, taxi stand etc wherever it is available. To counter the sensitivity of KR prediction performance due to influence of noisy GPS updates, the trips were simplified using RDP algorithm. For travel time prediction, additional features such as average speed, average acceleration and shape complexity (ratio of euclidean distance to haversine distance between end-points) were also considered. To speedup the process of feature extraction, an index structure based on geohash was used. Each GPS was represented using its geohash and then the nearest trips within the maximum distance threshold of 1km were searched using range queries.

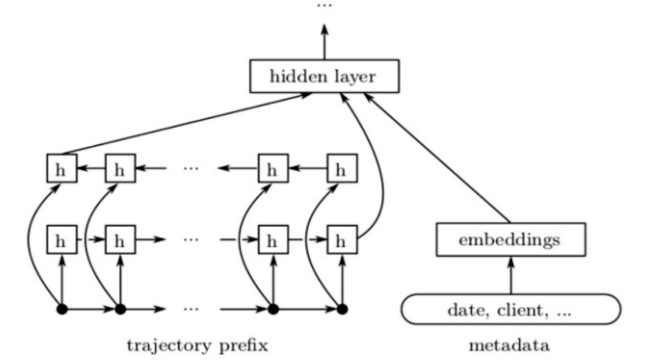


Fig. 3: Bidirectional RNN architecture.

Random Forest was used for Destination prediction. Support Vector Regression (SVR) also yielded similar results. For trip time prediction, ensemble of following regression models was used.

- Gradient Boosted Regression Trees (GBRT)
- Random Forest regressor (RF)
- Extremely Randomized Tree regressor (ERT) Stacked generalization approach was used to produce a single predictor. For destination prediction, Mean Haversine Distance(MHD) was used as an error measure of performance, where as Root Mean Square Logarithmic Error was used for the same in case of travel prediction problem.

C. Ensemble learning approach

The winner team for travel time prediction [3], used an ensemble learning approach. The framework here consists of hierarchy of expert models as given below

.

- Expert models for each test trip (e.g. trained on tracks which cross the test trip at the last known position).
- General base model: Based on a data set, where the features were extracted from all the tracks in the training set, and longer tracks were sampled more frequently than shorter ones.
- General expert models for short trips (e.g. only 1, 2 or 3positions of the initial trajectory are known).

The features used in this approach included haversine and cumulative distance of start and current positions from city center, median velocity, heading of the car in current position etc. i Either a random forest regression or a gradient boosting regression has been used was used as base classifier for ensemble modeling. Generally, bayesian optimization is use don a hold-out test set to tune weight factors. But, for this work, following heuristics were used –

- Use expert model for all test trips with sufficiently large training set.
- Use Average of all four models for all other test trips. Training sets were generated differently for different models as described below

.

- For base model, training set contained all trips.
- For expert model of short trips, a separate training set was built using first few GPS readings of all trips for each expert corresponding to trip length.
- For expert models for each test trip, a spatial clustering approach was used and all trips close to the current GPS position of taxi were selected. RMSLE was used as error measure for travel time prediction, whereas MHD was used for destination prediction. The key conclusion from this is that the remaining traveling time of a taxi depends mainly on the current position and heading of the taxi.

D. Trajectory distribution based approach

[2] uses a trajectory distribution based appraoch for destination prediction. This approach involves modeling of traffic flow pattern as a mixture of 2d guassian distributions. Known trajectories are then clustered using hierarchical clustering with ward-linkage criterion based on the Symetrized Segmentpath Distance. This distance compares trajectories as a whole, regardless of their time indexing or the number of locations that compose them. A new trajectory is then assigned to one of the clusters to predict the final destination. To predict the destination of new trajectory, only begining of its path is observed for a succession of locations. Contextual information such as hour of-the-day, day-of-the-week is considered in the prediction model using auxilary weights which is calculated as the product of any combination of three types of weights given below.

- Empiric Weight which describes the distribution information of traectory cluster

.

- Weekday weight which describes distribution information of the trajectory cluster at a given day of the week.
- Hours weight which describes the distribution of information of the trajectory cluster at a given hour of the day.

E. Tools and APIs for machine learning

There are number of tools which provide access to standard libraries of machine learning algorithms. [4] is one such popular tool which is being widely used for deep neural networks. R provides many libraries for statistical computing and analysis. These tools can be used for developing new models for this problem. Matlab, scilab, octave provide machine learning toolboxes. Weka is one another tool equipped with number of machine learning algorithms for classification and clustering. [1] used The a no, Block and Fuel for their implementation. Since, training phase is computationally expensive and takes longer, most of the times,

people have been using GPUs to train such models. With GPUs, the training time reduces considerably from few days to few weeks. [1] have used Nvidia GPUs for training the model

III. ACKNOWLEDGEMENT

My first and foremost acknowledgment is to my supervisor and guide Prof. Mr. A. M. Bagul During the long journey of this study, he supported mein every aspect. he was the one who helped and motivated me to propose research in this field and inspired me with his enthusiasm on research, his experience, and his lively character. I express true sense of gratitude to my guide Prof. Mr. A. M. Bagul for his perfect valuable guidance, all the time support and encouragement that he gave me. I would also like to thanks our head of department Mrs. S. A. Chiwhane, Principal Mr. R. S. Prasad, and management inspiring me and providing all lab and other facilities, which made this seminar presentation very convenient. I am really thankful to all those who rendered their valuable help for successful completion on seminar presentation.

REFERENCES

[1] A. de Br´ebisson, ´ E. Simon, A. Auvolat, P. Vincent, and Y. Bengio,"Artificial neural networks applied to taxi destination prediction," arXivpreprint arXiv:1508.00021, 2015.

[2] B. Guillouet, L. Jean-Michel, P. Besse, and R. Franc¸ois, "Destinationprediction by trajectory distribution based model," 2016.

[3] T. Hoch, "An ensemble learning approach for the kaggle taxi travel timeprediction challenge," ECML-PKDD-DCs, 2015.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick,S. Guadarrama, and T. Darrell, "Caffe: Convolutional archcture for fastfeature embedding," arXiv preprint arXiv:1408.5093, 2014.

[5] Kaggle, "Pkdd15 taxi trip time prediction," 2015. [Online]. Available: https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii

[6] H. T. Lam, E. Diaz-Aviles, A. Pascale, Y. Gkoufas, and B. Chen, "(blue)taxi destination and trip time prediction from partial trajectories," arXivpreprint arXiv:1509.05257, 2015.

[7] J. Mendes-Moreira and L. Moreira-Matias, "On learning from taxi-gps traces," ECML-PKDD-DCs, 2015.